W274–W276 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki493

SCANMOT: searching for similar sequences using a simultaneous scan of multiple sequence motifs

Saikat Chakrabarti, A. Prem Anand[1], Nitin Bhardwaj[2], Ganesan Pugalenthi and R. Sowdhamini*

National Centre for Biological Sciences (NCBS), Bangalore 560065, India, [1]International Institute of Information Technology (MSIT), Hyderabad 500 019, India and [2]Department of Chemical Engineering, Indian Institute of Technology, Mumbai 400 076, India

Received February 14, 2005; Revised and Accepted April 25, 2005

ABSTRACT

Establishment of similarities between proteins is very important for the study of the relationship between sequence, structure and function and for the analysis of evolutionary relationships. Motif-based search methods play a crucial role in establishing the connections between proteins that are particularly useful for distant relationships. This paper reports SCANMOT, a web-based server that searches for similarities between proteins by simultaneous matching of multiple motifs. SCANMOT searches for similar sequences in entire sequence databases using multiple conserved regions and utilizes inter-motif spacing as restraints. The SCANMOT server is available via http://www.ncbs.res.in/~faculty/mini/scanmot/scanmot.html.

INTRODUCTION

The most widely used tools for sequence similarity searching allow matching between arbitrary regions of the query and database sequences (1–5). Many motif-based search methods seek database sequences that match a pre-specified pattern (6–12). However, if this pattern is not specified with sufficient precision or is used singly, the number of matches can be very large, most of them having no biological relevance. On the other hand, a very specific pattern may exclude many sequences of interest.

Several programs like PHI-BLAST (13) also combine pattern matching with a search for statistically significant sequence similarity (10). But these approaches are limited to finding similar sequences on the basis of a single pattern at a time. Therefore, they do not utilize information embedded in multiple patterns for a given query protein.

In biological sequences, the occurrence of several conserved motifs is often more informative than the presence of a single motif. Multiple-motif-based search tools have been found to be useful in the past (MAST, Meta-MEME, 14,15). This paper describes a method which combines multiple-pattern searching with a search for statistically significant sequence similarity. The specificity of the search engine is increased by utilizing the inter-motif spacing and pairwise global alignment of the query and hits. A web-based server interface (SCANMOT), developed using this algorithm, is accessible via http://www.ncbs.res.in/~faculty/mini/scanmot/scanmot.html.

In addition, this server provides an option to validate PSI-BLAST (4) results and helps in attributing biological significance to the homologous sequences identified by statistical similarity. Therefore, this server can be of importance to the biological community for scanning and searching for distantly related sequences from different sequence databases.

METHODOLOGY AND SERVER FORMAT

SCANMOT works in three different steps to find similar proteins with the specified motifs.

(i) The first step is to scan the motifs within the query sequence itself to record the inter-motif spacing for all possible combinations of the specified motifs. The order of occurrence of the motif within the query sequence is also recorded.
(ii) Next, the program scans the motifs into a sequence database and reports the sequences with a specified number of motifs in a similar order to the query sequence.

*To whom correspondence should be addressed. Tel: +91 80 23636421/8, ext. 4240/1; Fax: +91 80 23636665; Email: mini@ncbs.res.in
Present addresses:
Saikat Chakrabarti, Computational Biology Branch, National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), NIH, 8600 Rockville Pike, Bethesda, MD 20894, USA
Prem Anand, Institute for Genomics and Bioinformatics, Graz University of Technology, Graz-8010, Austria
Nitin Bhardwaj, Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60612, USA

(iii) The third step is to align the query sequence and each of the 'hits' using an algorithm which fixes the motif regions between the query and hit and aligns the rest by dividing them into several parts (16).

All the sequences where the motifs are present are ranked by two independent grading methods, one is motif spacing and the other is alignment score. The SCANMOT server allows the user to select a relaxation filter for each residue within the motif for a wider range of 'hits' or similar sequences with allowed substitution.

Motif spacer

Different grades are assigned to the similar sequences ('hits') depending on the extent of the match in inter-motif spacing between the query and the hit.

For N motifs for a given query protein, $N \times (N - 1)/2$ combinations of inter-motif spacing are possible. For example, if a sequence has four motifs (A, B, C and D), six different motif-pair spacings are possible (AB, AC, AD, BC, BD and CD). Similarly, within the 'hit', if four motifs (a, b, c and d) are present there will be six different motif spacers (ab, ac, ad, bc, bd and cd). For each equivalent spacing between query and hit (e.g. AB and ab), a relaxation filter is applied to make the motif-based search more sensitive. Our benchmarking studies indicated that a 30% relaxation of equivalent motif spacing provides optimal results with high specificity. If the spacing between query and hit passes the relaxation filter, we select it as a matched motif pair (AB and ab). Grade A is assigned to a hit when the inter-motif spacing varies upto 20% compared with all possible combinations of motif spacers in the query sequence. Similarly Grades B, C and D are assigned when the inter-motif spacing varies between 20 and 40%, between 40 and 60% and by >60% compared with all possible motif spacer combinations in the query, respectively.

Percentile alignment score

The extent of similarity between the query and the hit is examined by alignment using a pairwise fixed motif alignment algorithm (16), where the alignment score is represented as a measure of similarity between the two. A percentile gradation scale is applied to all the query–hit alignments with respect to the top scoring alignment pair. In addition to the alignment score, hits are additionally validated using an amino acid similarity score and percentage sequence identity.

Proteins with known 3D structure are identified and also marked by the Protein Data Bank (PDB) (17,18) and SCOP (19) codes together with their accession number for better visualization and understanding of the scanning results. The frequency of occurrence of each motif is represented by a bar diagram and by labelling each motif with a different colour code.

Scanning similar sequences in custom databases

SCANMOT provides an option for searching similar sequences in well-curated sequence databases such as PDB, SCOP and SWISSPROT (20). In an example search of a potassium channel protein against a non-redundant sequence database, several hits are distantly related (∼85% of hits have <30% sequence identity to the query); interestingly, ∼60% of these distantly related hits retain different inter-motif spacing (Grade D), in contrast to closely related sequences, characterized by >90% sequence identity to the query, which are all Grade A hits (see the info pages on the web server for a histogram). This illustrates that it is possible to obtain more distantly related proteins by relaxing inter-motif spacing. There is also an option to search for similar motifs and sequences in individual genome databases of model organisms such as bacteria, yeast, worm, fly, mouse, human and plants. Further, the server allows the user to upload a custom defined sequence database of their choice and to search for similar motifs and sequences in that dataset.

Filtering and validating PSI-BLAST outputs

SCANMOT also provides options to filter PSI-BLAST (4) output for a given query sequence and allows the user to identify homologous sequences on the basis of the presence of motifs for a query protein family. This option extracts true homologous sequences defined in terms of the presence of motifs. The SCANMOT algorithm is employed to validate and characterize PSI-BLAST output to extract true homologues for protein families. SCANMOT is applied to every PSI-BLAST output, and homologous sequences are identified on the basis of the presence of motifs characteristic of the query family of proteins. The number of significant homologues identified by the SCANMOT motif scanning procedure is very close to the actual number of true positives (see Supplementary Material for the results for nine families).

RESULTS AND APPLICATIONS

SCANMOT has been benchmarked using proteins that are related at the family and superfamily levels (see Supplementary Material for details). Large-scale benchmarking studies using structural motifs for 110 superfamilies from the SMoS database (21) yielded very high specificity (86%) and 70% coverage. New connections could also be obtained with several 'hypothetical' entries in the sequence database (22; see Supplementary Material). SCANMOT can be a useful tool for the identification of distant relationships among proteins. Large-scale genome-wide surveys for specific sequences and motifs can utilize SCANMOT as a rapid validation tool. Most genome-wide surveys require the examination of sequence motifs to confirm relationships and assign functional information to hypothetical proteins.

CONCLUSION

SCANMOT is a server that utilizes the occurrence and position of several conserved motifs along a protein sequence. It is possible to search for similar sequences in entire sequence databases using these conserved regions and the spacing as sole restraints. The utilization of multiple motifs during the scanning procedure can drastically reduce the rate of false positives and at the same time extract novel true positives. Therefore, careful utilization and examination of such results can provide useful information for future motif-based research. SCANMOT can also be useful for such applications as the investigation of distant relationships and cross-family evolutionary connections among proteins.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

R.S. is a Senior Research Fellow funded by the Wellcome Trust Grant. G.P. is supported by the Wellcome Trust. We would also like to acknowledge the National Centre for Biological Sciences (TIFR) for financial and infrastructural support. The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
2. Pearson,W.R. and Lipman,D.J. (1988) Abstract improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
3. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
4. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
5. Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol.*, **266**, 460–480.
6. Myers,E.W. and Miller,W. (1989) Approximate matching of regular expressions. *Bull. Math. Biol.*, **51**, 5–37.
7. Smith,R.F. and Smith,T.F. (1990) Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc. Natl Acad. Sci. USA*, **87**, 118–122.
8. Staden,R. (1990) Searching for patterns in protein and nucleic acid sequences. *Methods Enzymol.*, **183**, 193–211.
9. Mehldau,G. and Myers,G. (1993) A system for pattern matching applications on biosequences. *Comput. Appl. Biosci.*, **9**, 299–314.
10. Tatusov,R.L. and Koonin,E.V. (1994) A simple tool to search for sequence motifs that are conserved in BLAST outputs. *Comput. Appl. Biosci.*, **10**, 457–459.
11. Ogiwara,A., Uchiyama,I., Takagi,T. and Kanehisa,M. (1996) Construction and analysis of a profile library characterizing groups of structurally known proteins. *Protein Sci.*, **5**, 1991–1999.
12. Bairoch,A., Bucher,P. and Hofmann,K. (1997) The PROSITE database, its status in 1997. *Nucleic Acids Res.*, **25**, 217–221.
13. Zhang,Z., Schaffer,A.A., Miller,W., Madden,T.L., Lipman,D.J., Koonin,E.V. and Altschul,S.F. (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.*, **26**, 3986–3990.
14. Bailey,T.L. and Gribskov,M. (1997) Score distributions for simultaneous matching to multiple motifs. *J Comput Biol.*, **4**, 45–59.
15. Grundy,W.N., Bailey,T.L., Elkan,C.P. and Baker,M.E. (1997) Meta-MEME: motif-based hidden Markov models of biological sequences. *Comput. Appl. Biosci.*, **13**, 397–406.
16. Chakrabarti,S., Bhardwaj,N., Anand,P.A. and Sowdhamini,R. (2004) Improvement of alignment accuracy utilizing sequentially conserved motifs. *BMC Bioinformatics.*, **5**, 167–179.
17. Bernstein,F.C., Koetzle,T.F., Williams,G.J., Meyer,E.F.,Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur. J. Biochem.*, **80**, 319–324.
18. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
19. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
20. Apweiler,R., Gateau,A., Contrino,S., Martin,M.J., Junker,V., O'Donovan,C., Lang,F., Mitaritonna,N., Kappus,S. and Bairoch,A. (1997) Protein sequence annotation in the genome era: the annotation concept of SWISS-PROT, TREMBL. In *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology (ISMB-97)*, 21–25 June, Halkidiki, Greece. AAAI Press, Menlo Park, pp. 33–43.
21. Chakrabarti,S., Venkatramanan,K. and Sowdhamini,R. (2003) SMoS: a database of structural motifs of protein superfamilies. *Protein Eng.*, **16**, 791–793.
22. Chakrabarti,S. and Sowdhamini,R. (2004) Regions of minimal structural variation among members of protein domain superfamilies: application to remote homology detection and modelling using distant relationships. *FEBS lett.*, **569**, 31–36.